Econometrics and Statistics 000 (2016) 1–9

Contents lists available at ScienceDirect

Econometrics and Statistics

journal homepage: www.elsevier.com/locate/ecosta

A Fisher-scoring algorithm for fitting latent class models with individual covariates

Antonio Forcina*

Dipartimento di Economia, via Pascoli, Perugia 06100, Italy

ARTICLE INFO

Article history:
Received 27 February 2016
Revised 18 July 2016
Accepted 18 July 2016
Available online xxx

Keywords:
Categorical data analysis
EM algorithms
Empirical information matrix
Fisher scoring algorithms
Individual covariates
Latent class models
Line search
multinomial logit

ABSTRACT

Describes a modified Fisher scoring algorithm for fitting a wide variety of latent class models for categorical responses when both the class weights and the conditional distributions of the responses depend on individual covariates through a multinomial logit. A simple expression for computing the score vector and the empirical information matrix is presented; it is shown that this matrix is positive definite under mild conditions. The Fisher scoring algorithm combines the empirical information matrix to update the step direction with a line search to optimize the step length. The algorithm converges for almost any choice of starting values. An application to the field of education transmission seems to suggest that, while parents' education affects the child latent ability, their pressure affects directly the child's achievements.

1. Introduction

Latent class models are important tools for modeling the dependence structure among categorical responses induced by unobservable heterogeneity. Since the pioneering work of Goodman (1974), extensions in several directions have been proposed. Particular attention has been devoted to the case where the probabilities of belonging to the different latent classes may depend on covariates, see, for a detailed discussion, Vermunt (2010) and Petersen et al. (2012). Latent class models where covariates may also affect the distribution of the responses conditionally on the latent, have been used by Bartolucci and Forcina (2006) and Dardanoni and Li Donni (2012) among others. For a general discussion of identifiability and inference in latent class models where covariates may affect both the marginal distribution of the latent and the conditional distribution of the responses, see Huang and Bandeen-Roche (2004). Finally, models where the assumption of conditional independence may be violated have been considered, among others, by Hagenaars (1988) within a log-linear context and Bartolucci and Forcina (2006) in the context of marginal models.

It may be useful to recall the main reasons for using the EM (expectation–maximization) algorithm to compute the maximum likelihood estimates in mixture models: (i) the likelihood of the complete data is, usually, much simpler to maximize, (ii) the algorithm is numerically stable, (iii) the likelihood of the incomplete data always increases from an M step to the next. One drawback of the EM algorithm is that, sometimes, the improvement produced by a single step may be so small that it may even becomes difficult to determine when to stop. In the case when covariates are assumed to affect only

* Fax: +390755855950.
E-mail addresses: forcina@stat.unipg.it, forcinarosara@gmail.com

http://dx.doi.org/10.1016/j.ecosta.2016.07.001

Please cite this article as: A. Forcina, A Fisher-scoring algorithm for fitting latent class models with individual covariates, Econometrics and Statistics (2016), http://dx.doi.org/10.1016/j.ecosta.2016.07.001

the latent weights, methods for computing approximate estimates have been proposed by Bolck et al. (2004) and Vermunt (2010). To speed up convergence, hybrid algorithms which combine EM and quasi-Newton steps have been proposed, see for instance (Lin and Lee, 2008). Most available packages combine the EM algorithm with some kind of quasi-Newton step to increase efficiency. The performance of quasi-Newton algorithms may be greatly improved by performing a suitable line search to optimize the step length (see for example Potra and Shi, 1995; Turner, 2008).

A Fisher scoring algorithm is described for fitting a rather general family of latent class models where individual covariates may affect both the marginal distribution of the latent and the conditional distributions of the responses. In addition, response variables are allowed to be associated conditionally on the latent, as long as the model is identifiable. Though this class of models is not entirely new and not as general as the one described by Bartolucci and Forcina (2006), its likelihood is considerably easier to handle. By introducing a convenient matrix notation, a simple expression for the score vector and the *Empirical* information matrix (EIM) Meilison (1989) is derived; this matrix is used for computing the step direction. The EIM is much simpler to compute than both the observed and the expected information matrices. It is shown that the EIM, like the expected information matrix, is positive definite under very mild conditions, a result which does not hold for the observed information matrix. The use of the EIM for fitting algorithms and for estimating standard errors has been investigated by Scott (2002) in a quite general context.

Section 2 defines the model, derives an expression for the score and the empirical information matrix and gives conditions under which this matrix is positive definite. Section 3 discusses efficient computation of these quantities and describes a line search algorithm which makes the proposed Fisher scoring algorithm efficient and stable at the same time. Section 4 contains an application to the field of education transmission; Section 5 is devoted to discussions and main conclusions.

2. Description of the model and main results

In the following let $\mathbf{1}_u$ and $\boldsymbol{I}_v$ denote, respectively, a column vector of u ones and an identity matrix of size v. Suppose that there are n subjects and c latent classes with the latent variable U coded as $1, \dots, c$. Let $\boldsymbol{\pi}_i$, $i = 1, \dots, n$, denote the vector whose elements are the probabilities that the ith subject belongs to one of the c latent classes. Suppose there are T categorical response variables with V_t, $t = 1, \dots T$, having r_t categories; let $r = \prod_t r_t$ be the number of configurations of the response variables. The joint distribution of the response variables for the ith subject, conditional on the latent $U = j$, $j = 1, \dots, c$, is determined by the $r \times 1$ vector of probabilities $\boldsymbol{q}_{ij}$ whose entries are arranged in lexicographic order, meaning that response variables with a larger index run faster. Finally let $\boldsymbol{p}_i = \Sigma_j \pi_{ij} \boldsymbol{q}_{ij}$ denote the $r \times 1$ vector of probabilities for the joint distribution of the responses.

2.1. A model for the latent distribution

Let $\boldsymbol{X}_i$ be a $c \times K$ matrix of known constants, possibly depending on individual covariates, $\boldsymbol{\beta}$ a vector of K unknown parameters and assume that the marginal distribution of the latent is determined by a multinomial logit model with the initial category as reference,

$$\boldsymbol{\pi}_i = \frac{\exp(\boldsymbol{X}_i \boldsymbol{\beta})}{\mathbf{1}_c' \exp(\boldsymbol{X}_i \boldsymbol{\beta})}.$$

When no covariates are available, $\boldsymbol{X}_i$ will be an identity matrix of size c without the first column. To make the jth logit ($j = 2, \dots, c$) a linear function of x_i, simply add a column of zeros except for the jth element which is equal to x_i.

2.2. A model for the conditional distributions of the responses

The formulation proposed below is, essentially, a log-linear model with individual covariates for the $\boldsymbol{q}_{ij}$, $j = 1, \dots, c$, the vectors containing the joint distribution of the responses for the ith subject conditionally on the latent. It is convenient to define the model in three steps as detailed below.

(i) Set up a log-linear model which can be expressed in the form of a multivariate logit

$$\boldsymbol{q}_{ij} = \frac{\exp(\boldsymbol{G}\boldsymbol{\theta}_{ij})}{\mathbf{1}_r' \exp(\boldsymbol{G}\boldsymbol{\theta}_{ij})},$$

where $\boldsymbol{G}$ is a $r \times g$ full rank design matrix for the log-linear model with $r \gg g$ and $\boldsymbol{\theta}_{ij}$ ($j = 1, \dots, c$) is a vector of log-linear parameters possibly depending on covariates. If responses are assumed to be conditionally independent, $\boldsymbol{G}$ includes only the main effects of the response variables. Identifiability of models when interactions terms are present may be problematic (see Stanghellini and Vantaggi, 2013). An example of a design matrix $\boldsymbol{G}$ with interaction terms between responses coded in a parsimonious way is given in Section 4. An algorithm for constructing the $\boldsymbol{G}$ matrix for any given log-linear model under the reference category coding is given by Colombi and Forcina (2014), Section 2.1.

Please cite this article as: A. Forcina, A Fisher-scoring algorithm for fitting latent class models with individual covariates, Econometrics and Statistics (2016), http://dx.doi.org/10.1016/j.ecosta.2016.07.001